Abstract

This report is committed to build a model which is used to determine the relationship between several kinds of given data from Share price and other daily trading data about Bank of Ningbo, especially focus on the 'ClosePrice' (The closing price of the share on the current day) and other corresponding parameters. It had been found out that the 'ClosePrice' was strongly related to 2 different kinds of given data which will be shown in this report. Finally, this report used the built model to predict and analysis the future data.

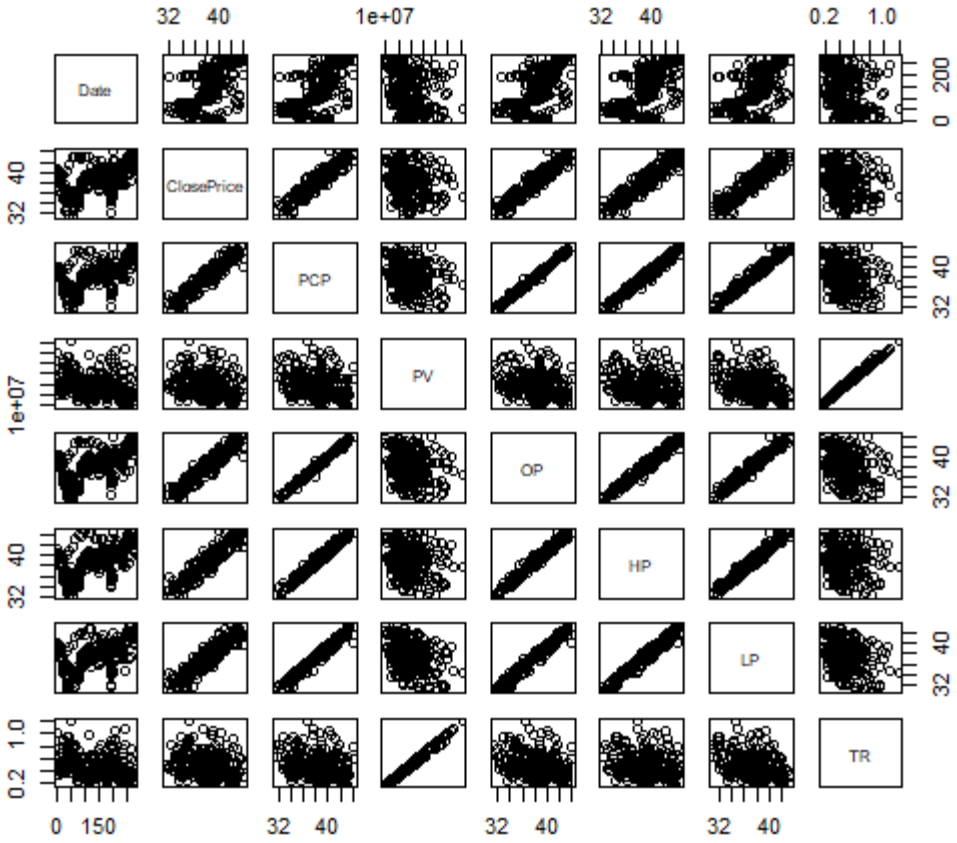
Introduction

This report is focusing on using R studio to solve the given tasks ---- finding the association between 'ClosePrice' and several parameters (PCP; PV; OP; HP; LP; TR)(1). The data sheets used in this report (NBCBtrain.txt & NBCBtest.txt) were given from SMM moodle page. The coding used in this report will also be given. Please turn off the 'automatic error correction' function of Word to make sure the highlights could be seen clearly.

Modeling Process

First, the model was needed to be constructed. 2 useful packages were needed to be downloaded: 'dplyr' and 'leaps' (Please download the package first to make sure the code could be run). Data from data sheets was imported and visualized. Please notice the data should be imported from a right source.

```
NBCBtrain <- read.delim("C:/Users/ssyxy3/Desktop/NBCBtrain.txt")
data(NBCBtrain)
str(NBCBtrain)

library(dplyr)
library(leaps)
```



The data was selected by choosing the random variable ‘Data’.

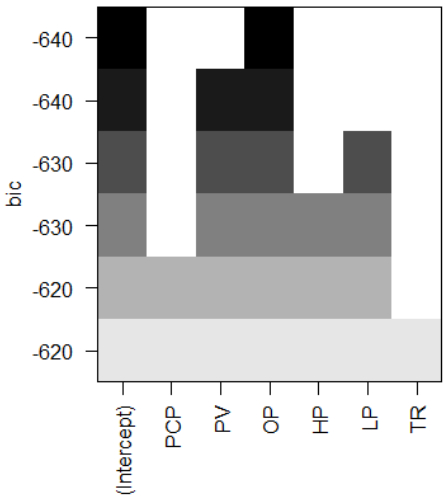
```
Train<-select(NBCBtrain, -Date)
Test<-select(NBCBtrain, -Date)
TestResponses = select(Test, ClosePrice)$ClosePrice

Subset selection object
Call: regsubsets.formula(ClosePrice ~ ., data = Train)
6 Variables  (and intercept)
    Forced in Forced out
PCP     FALSE      FALSE
PV      FALSE      FALSE
OP      FALSE      FALSE
```

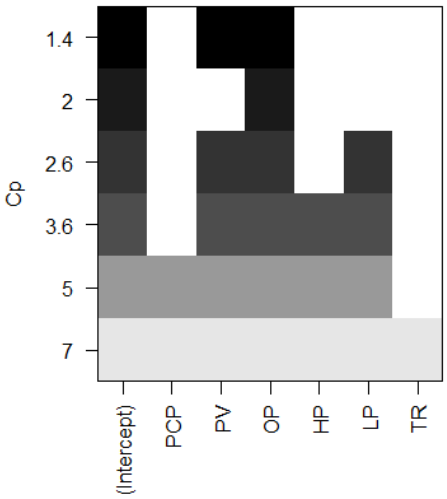
```
HP     FALSE     FALSE
LP     FALSE     FALSE
TR     FALSE     FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
                          PCP         PV          OP          HP
LP          TR
1       (     1     )   "    "    "    "    " * "    "    "    "    "    "    "
2       (     1     )   "    "    " * "    " * "    "    "    "    "    "    "
3       (     1     )   "    "    " * "    " * "    "    "    " * "    "    "
4       (     1     )   "    "    " * "    " * "    " * "    " * "    "    "
5       (     1     )   " * "    " * "    " * "    " * "    " * "    " * "    "
6       (     1     )   " * "    " * "    " * "    " * "    " * "    " * " * "
```

Next it was needed to decide which models will be used by comparing the value of $MSE^{(2)}$ of different models. The smaller value of MSE the model had, the better accuracy the model had.

```
predictions<-predict(fit, newdata=select(Test, -ClosePrice))
mse_MSE<-mean((predictions-TestResponses)^2)
mse_MSE
```

By using the method of Best subsets regression, 3 models were suggested. The reason of choosing this method was provided below.






```
fit1<-lm(ClosePrice ~ OP, data=Train)
fit2<-lm(ClosePrice ~ OP + PV, data=Train)
fit3<-lm(ClosePrice ~ OP + PV + LP, data=Train)

predictions<-predict(fit1, newdata=select(Test, -ClosePrice))
mse_bestbets1<-mean((predictions-TestResponses)^2)

predictions<-predict(fit2, newdata=select(Test, -ClosePrice))
mse_bestbets2<-mean((predictions-TestResponses)^2)

predictions<-predict(fit3, newdata=select(Test, -ClosePrice))
mse_bestbets3<-mean((predictions-TestResponses)^2)

mse_bestbets1
mse_bestbets2
mse_bestbets3
```

The ‘best bets’ were constructed and selected based on the method of Best Subsets Regression which was focusing on the BIC[3]. In fact, 2 kinds of method were provided: Best

Subsets Regression and Stepwise Regression. The first method was chosen since there were 6 random variables at most. The more detailed information was given by the first method when the number of random variables was small. Besides, the Best Subsets Regression method provided different kinds of models to be chosen, which also provided more flexible ways to analysis the data.

The results of MSE computation was given below.

mse_bestbets1: 0.7345048

mse_bestbets2: 0.7275959

mse_bestbets3: 0.7256482

mse_MSE (full model): 0.7213426

The different models' MSE were computed, it was simple to find out the third model had the best accuracy since it had the closest value to the value of MSE of the full model. However, the first model had more concise construction which had only 1 random variable, less than the third model. Moreover, the first model also had very close value of MSE to the full model. After tradeoff the accuracy and complexity, the first model was chosen to be the 'best model'.

Results

After determining the best model, the best model was used to predict the future data of 'ClosePrice' based on the other corresponding parameters. The coefficient of the chosen model was computed.

lm(formula = ClosePrice ~ OP, data = NBCBtrain)

Coefficients:

(Intercept)	OP
1.9801	0.9495

ClosePrice* = *Intercept* + *bOP* where *b* = 0.9495

This formula was used to compute the future data of ClosePrice. The computation result was compared to the actual value which was given. The data was listed below.

Date	(Actual) ClosePrice	Model 1
Day281	36.1	36.42796
Day282	35.89	35.87725
Day283	36.81	36.06715
Day284	37.09	37.20655
Day285	37.39	37.08312
Day286	38.5	37.27302
Day287	38.93	37.97565
Day288	38.75	38.94414
Day289	39.67	39.18151
Day290	39.7	39.45687
Day291	38.58	39.18151
Day292	38.79	38.43141
Day293	37.57	38.91565
Day294	38.31	37.58635

The absolute difference between actual value and the predictions was calculated. The mean value was calculated by Excel or R studio.

**mean absolute error = [sum of (| actual value – prediction value of model 1 |)]/n*

= 0.545964

This value will be used to determine whether the chosen model could work successfully in an acceptable error range. In conclusion, the first model was determined to compute the future data in this part, and the mean absolute error between the actual value and the predictions was computed.

Exploratory Analysis

To make exploratory analysis, the chosen model was compared to the full model which had been mentioned above. The comparation was used to determine the difference of error between two different models and actual value.

lm(formula = ClosePrice ~ PCP + PV + OP + HP + LP + TR, data = NBCBtrain)

Coefficients:

(Intercept)	PCP	PV	OP	HP	LP	TR
1.081e+00	-1.567e-01	1.474e-08	1.042e+00	-1.320e-01	2.151e-01	1.099e-02

*$ClosePrice = intercept + a_1PCP + a_2PV + a_3OP + a_4HP + a_5LP + a_6TR$

where $a_1 = -1.57*10^{-1}$; $a_2 = 1.47*10^{-8}$; $a_3 = 1.04$ $a_4 = -1.32*10^{-1}$; $a_5 = 2.15*10^{-1}$; $a_6 = 1.10*10^{-2}$

This formula was used to compute a different prediction of ClosePrice, which was given below.

Date	**(Actual) ClosePrice**	Full Model
Day281	36.1	36.31082
Day282	35.89	35.75854
Day283	36.81	35.8895
Day284	37.09	37.10987
Day285	37.39	37.02014
Day286	38.5	37.147
Day287	38.93	37.76308
Day288	38.75	38.80343
Day289	39.67	39.20129
Day290	39.7	39.49781
Day291	38.58	39.11352
Day292	38.79	38.44275
Day293	37.57	38.91518
Day294	38.31	37.63622

The formula given above was used to compute the mean absolute error again.

*$mean\ absolute\ error = [\ sum\ of\ (\ |\ actual\ value - prediction\ value\ of\ full\ model\ |\ )\ ]/n$

$= 0.556891$

Hence it was found that the difference of error between the chosen model and the full model is quite small, which means that the accuracy of the chosen model is high.

Summary and Conclusion

This report was focused on finding the relationship between the ClosingPrice and other several corresponding parameters of NBCB. By constructing a linear model, it was found that the ClosingPrice had the strongest linearly correlation with the OpeningPrice. Besides, the ClosingPrice had a negative linearly correlation with PCP and HP, it had a quite weak linearly correlation with PV. It was shown that if the ClosingPrice was needed to predict, the OpeningPrice would be the most influential parameter.

The chosen model still can be optimized by considering the other kinds of relationship between the ClosingPrice and given parameters but not linearly correlation. This idea was not concluded in the content of this report.

Appendix 1: the explanation of superscripts and notation

(1):

ClosePrice: The closing price of the share on the current day.

PCP: The closing price of the share on the previous day.

PV: The volume of the share traded on the previous day.

OP: The opening price of the share on the current day.

HP: The highest price of the share on the previous day.

LP: The lowest price of the share on the previous day.

TR: The rate of turnover (%) of the share on the previous day.

(2): Mean Square Error (MSE)

(3): Bayesian Information Criterion (BIC)

*: Formula based on models and basic statistics computation

Appendix 2: Code of R with annotations

```
#Here we go
#First import the data, and check the data from 2 txt
NBCBtrain <- read.delim("C:/Users/ssyxy3/Desktop/NBCBtrain.txt")
data(NBCBtrain)
str(NBCBtrain)

#We need to select the data and compute the MSE
library(dplyr)#downloaded
Train<-select(NBCBtrain, -Date)
Test<-select(NBCBtrain, -Date)
TestResponses = select(Test, ClosePrice)$ClosePrice
fit<-lm(ClosePrice~., data=Train)
coef(fit)
predictions<-predict(fit, newdata=select(Test, -ClosePrice))
mse_MSE<-mean((predictions-TestResponses)^2)
mse_MSE

#Now we need to decide the best model we should use
library(leaps)#downloaded
a<-regsubsets(ClosePrice~., data=Train)
summary.out<-summary(a)
summary.out
summary.out$cp
plot(a, scale='Cp')
plot(a, scale='bic')

#If we focus on BIC, we should try 3 models
fit1<-lm(ClosePrice ~ OP, data=Train)
fit2<-lm(ClosePrice ~ OP + PV, data=Train)
fit3<-lm(ClosePrice ~ OP + PV + LP, data=Train)

predictions<-predict(fit1, newdata=select(Test, -ClosePrice))
mse_bestbets1<-mean((predictions-TestResponses)^2)
```

```
predictions<-predict(fit2, newdata=select(Test, -ClosePrice))
mse_bestbets2<-mean((predictions-TestResponses)^2)

predictions<-predict(fit3, newdata=select(Test, -ClosePrice))
mse_bestbets3<-mean((predictions-TestResponses)^2)

mse_bestbets1
mse_bestbets2
mse_bestbets3

lm(formula = ClosePrice ~ OP, data = NBCBtrain)
lm(formula = ClosePrice ~ PCP + PV + OP + HP + LP + TR, data = NBCBtrain)
```